基于熵和 SVM 多分类器的异常流量检测方法

朱佳佳,陈 佳
(北京交通大学 电子信息工程学院,北京 100044)

摘 要:随着大数据时代的到来,各种数据挖掘和机器学习方法被广泛地应用于异常流量检测。文中针对异常流量检测方法展开研究,提出了一种基于熵和改进的 SVM 多分类器的异常流量检测方法。该方法用熵值对网络流量的各个属性进行量化,将异常流量检测问题抽象为对不同类型流量的分类问题,并对传统的一对其余 SVM 多分类器进行改进。使用改进 SVM 多分类器对熵值量化后的流量进行分类判决,根据分类结果捕获异常。将该方法应用于实际的异常流量检测系统,并进行测试,结果表明,该方法对网络中常见的异常流量有很好的检测效果。

关键词:异常检测;信息熵;一对其余;分类
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)03-0031-05
doi:10.3969/j.issn.1673-629X.2016.03.008

An Anomaly Detection Method Based on Entropy and SVM Multi-class Classifier

ZHU Jia-jia, CHEN Jia
(School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With the advent of the age of big data, data mining and machine learning methods have gradually replaced the traditional methods of anomaly detection, which have gained more attention. In this paper, a new method of detecting the anomaly traffic based on the information entropy and SVM is proposed. This method transfers anomaly detection problems into the classification of different types of traffic, and uses information entropy to quantify different attributes of network traffic. It puts forward an improved SVM multi-class classifier to classify the entropy-quantified traffic and judges the anomalies accordingly. This method is implemented into a real system and function test is carried out. The results show that the method has a good detection effect for the abnormal traffic of the Internet.

Key words: anomaly detection; information entropy; one-to-all; classification

0 引 言

异常流量检测是网络安全监管系统的重要组成部分,它通过监测和分析网络流量,发现和判别网络中的异常行为,帮助网络管理员及时采取有效的措施填补系统漏洞,保障系统安全。

在异常流量检测方法的相关研究中,网络流量的特征量化问题一直广受关注[1-3]。香农将熵的概念引入信息论中,利用熵值度量随机事件的不确定性。相关研究表明[4-6],正常流量与异常流量的不同属性在分布特征上存在明显差异。因此,可以通过分析网络流量特征信息的熵值变化规律来检测异常,从而解决网络流量的特征量化问题。

在异常流量检测系统中,需要一定数量的训练样本来建立正常或异常流量模型[7]。但是,在实际网络环境中,获得所需的数据并不容易,采集得到的数据源往往具有维数高、样本数小等问题[8]。与其他分类算法相比,SVM 可以更好地解决小样本、非线性、高维数等问题,因此非常适合被应用于异常检测系统中[9]。

文中结合信息熵理论与 SVM 多分类算法,将异常流量检测问题抽象为对不同类型流量的分类问题,用 SVM 多分类器对熵值量化后的流量进行分类判决,该方法具有检测精度高和检测速度快等优点。

收稿日期:2015-06-12 **修回日期**:2015-09-15 **网络出版时间**:2016-02-18
基金项目:国家重大专项(2013ZX03006002);国家自然科学基金资助项目(61471029);北京市自然基金“面上”项目(4132053);基本科研业务费(2014JBM012)
作者简介:朱佳佳(1990-),女,硕士,研究方向为信息网络关键技术;陈 佳,博士,副教授,研究方向为新一代信息网络理论与关键技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160218.1634.042.html

1 信息熵和流量的熵值量化

1.1 信息熵的定义

在信息论中,熵是对不确定性的一种度量。对于一条消息来说,熵值越高,说明该消息包含的信息量越大,反之说明该消息包含的信息量越小[10]。

信息熵的定义如下:

$$H(X) = -\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\ln\left(\frac{n_i}{S}\right)$$

其中:$X = \{n_i, i = 1,2,\cdots,N\}$ 表示在测量数据中属性 i 发生了 n_i 次;$S = \sum_{i=1}^{N} n_i$ 表示某个属性发生的总次数。

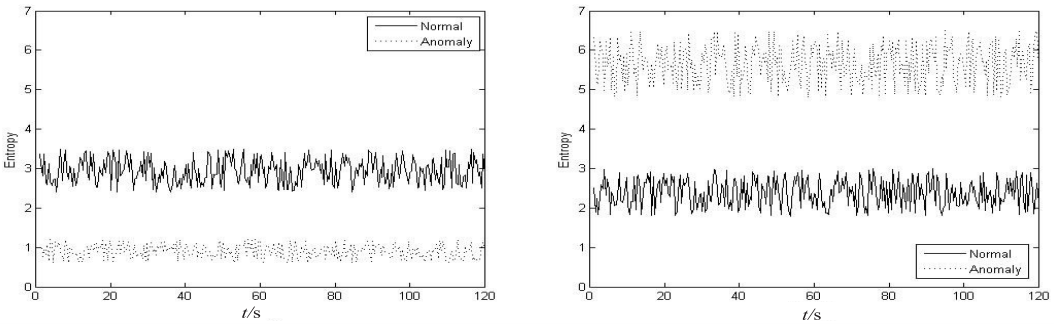
1.2 流量的熵值量化方法

众多研究表明,网络流量中的 IP 地址、协议和端口等属性在分布特征上表现出较强的自相似特性和重尾特性[11],正常流量和异常流量的各个属性在分布特征上存在明显的差异。因此,可以将信息熵应用于异常检测,用熵值量化网络流量的不同属性。

将捕获到的数据包按照时间顺序以每 1 000 个数据包为单位划分为一个子集,并规定其为单位流量,记为 $S_i = \{s_1, s_2, \cdots, s_{1\,000}\}$。然后选取源 IP、目的 IP、协议类型、源端口和目的端口等属性,分别计算单位流量在这些属性上的熵值。

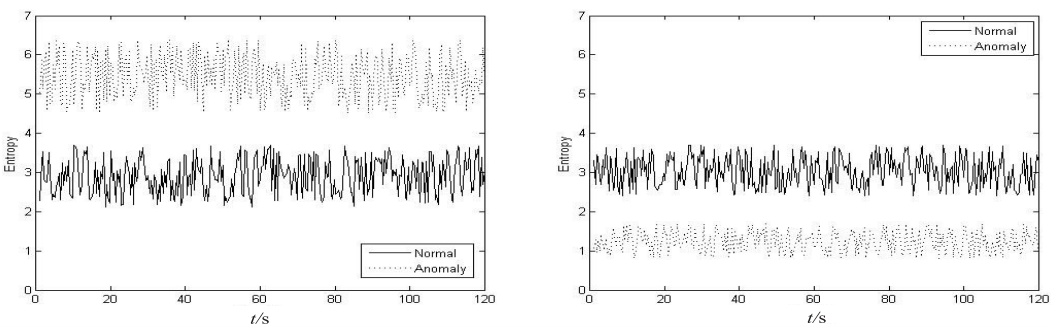
以源 IP 为例,N 为单位流量 S 中出现的不同的源 IP 的个数,$n_i(i = 1,2,\cdots,N)$ 为不同的源 IP 分别出现的次数,则 $S = \sum_{i=1}^{N} n_i = 1\ 000$,代入信息熵的定义公式即可得到单位流量源 IP 的熵值。

为了验证信息熵对正常流量和异常流量在特征差异上的表达能力,通过实验对正常流量、端口扫描异常流量和 DDOS 攻击异常流量在不同属性上的熵值进行对比。图 1(a)、(b)分别是正常流量端口扫描异常流量在目的 IP 和目的端口的熵值曲线;图 2(a)、(b)分别是正常流量与 DDOS 攻击异常流量在源 IP 和目的 IP 上的熵值曲线。



(a)目的 IP 的熵值曲线 (b)目的端口的熵值曲线

图 1 正常流量与端口扫描异常流量熵值对比



(a)源 IP 熵值曲线 (b)目的 IP 熵值曲线

图 2 正常流量与 DDOS 攻击异常流量熵值对比

相对于正常流量,端口扫描异常流量的目的 IP 更集中,目的端口更分散,所以目的 IP 熵值较小,目的端口熵值较大;DDOS 攻击会使用大量的僵尸主机,或产生大量虚假源 IP 地址实施攻击,所以 DDOS 攻击异常流量的目的 IP 熵值较小,源 IP 熵值较大。

鉴于正常流量和异常流量的不同属性在分布特征上的明显差异,可以将异常流量的检测问题转换为一个基于流量属性熵值的分类问题。

2 SVM 和异常流量检测

2.1 改进的 SVM 多分类算法

目前,基于二分类的 SVM 多类分类方法主要有“一对一”、“一对其余”、DAG-SVM 和输出纠错编码法等几种方法[12-14]。“一对其余”是现在应用较为广泛的多类分类方法。对于“一对其余”算法,假设有 k 种样本,则需要构造 k 个二类分类器,每一个分类器用

于把其中一类与其余各类分开。训练时,取其中一类为正类,其余 $k-1$ 类为负类。判决时,待检测样本顺序经过 k 个二类分类器共得到 k 个输出值 $f_i(x)=\mathrm{sgn}(g_i(x))$,$i=1,2,\cdots,k$。若判决结果中仅包含一个+1,则待检测样本类别为对应分类器的正类类别。若判决结果中存在不只一个+1,即出现分类重叠现象,还需比较输出为+1 的分类器的决策函数值,值最大的分类器的正类代表待检测样本的类别。若判决结果都是-1,则认为该样本不可分。

鉴于"一对其余"方法可能遇到的分类重叠和不可分问题,文中提出一种改进的一对其余 SVM 多分类算法。该方法利用聚类分析中的类距离思想作为检测模型中各个二类分类器排列顺序的依据。对于 k 类样本,首先计算每一类到其他各类的中心距离,然后计算每一类到其他各类的平均距离。平均距离最大的类是特异性最明显的类,优先选取这样的类作为排列靠前的二类分类器的正类。距离的相关定义如下:

定义 1:中心距离。

第 i 类和第 j 类样本的中心距离定义为空间中能包含所有第 i 类样本的球面中心到能包含所有第 j 类样本的球面中心的欧式距离,记为 d_{ij};

定义 2:平均距离。

第 i 类到其余各类的平均距离定义为第 i 类样本到其他各类样本中心距离的平均值,记为 γ_i,且满足 $\gamma_i=\left(\frac{1}{k-1}\right)\sum_{j=1}^{k}d_{ij}(i\neq j)$。

具体步骤如下:

步骤 1:根据定义 1 计算各类到其他类的中心距离 $d_{ij}(i,j=1,2,\cdots,k,i\neq j)$;

步骤 2:根据定义 2 计算各类到其他类的平均距离 $\gamma_i(i=1,2,\cdots,k)$;

步骤 3:比较步骤 2 中 γ_i 的大小,然后按照 γ_i 从大到小的顺序为各个类别编号;

步骤 4:按照步骤 3 中得到的样本编号顺序逐个构造 k 个二类分类器。编号排列第一的样本作为第一个二类分类器的正类,编号排列第二的样本作为第二个二类分类器的正类,以此类推。

检测时,让待测样本依次通过各个二类分类器,如果待测样本在某一个分类器中的判决结果为+1,则判定该样本为对应此分类器的正类的类别,并终止此样本的检测。若样本依次经过所有的二类分类器输出结果都为-1,则判定为未知流量,加入待验证集,等待重新训练。样本判决过程如图 3 所示。

与传统的"一对其余"方法相比,距离优先 SVM 多分类算法可以使越靠前的分类器检测精度越高。因此,虽然在建立模型集时都构造了 k 个二类分类器,但是改进 SVM 一对其余多分类算法并不需要所有的样本都经过 k 个分类器,而是一旦被某个分类器判决为+1,则终止判断,有效缩短了样本的检测时间。

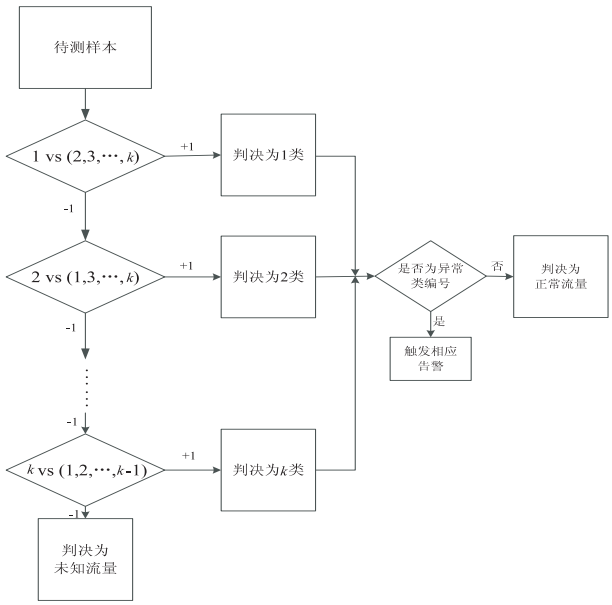


图 3　改进的"一对其余"SVM 多类分类器决策算法示意

2.2　SVM 分类思想

SVM 是统计学中以最小化结构风险为原则的一种新型机器学习方法。对于简单的二值分类问题,假设某空间中的样本可以表示为 (x_i,y_i),其中 x_i 为以向量表示的第 i 个样本,y_i 代表样本的分类标记,在二值分类问题中,y_i 的取值是 1 或-1。SVM 要解决的问题就是在空间中找到一个决策平面 $g(x)=wx+b$,使标记为 1 和-1 的样本分别位于决策平面的两侧[15]。对于待测样本,只要知道它位于决策平面的哪侧,就可以确定该样本的分类标记。

2.3　异常流量检测方法

按照 1.2 节提到的方法对所有捕获到的数据包进行单位流量划分,计算各个单位流量五元组的熵值,将每个单位流量表示为一个包含 5 个熵值的向量,记为:

$$\boldsymbol{V}_i=(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5)^{\mathrm{T}}$$

其中:α_1 到 α_5 分别代表源 IP、目的 IP、协议类型、源端口和目的端口的熵值;$\boldsymbol{V}_i$ 为熵值向量。

众多的熵值向量构成了 SVM 检测的样本集,其中已知类型的样本通过类型标记可以用来构造训练集并通过训练生成检测模型,而未知类型的样本则构成检测集。具体的检测方法如下:

(1)构造包含 N 种流量的样本集,包括正常流量和 $N-1$ 种在流量分布特征上差异较明显的异常流量;

(2)采用改进的一对其余 SVM 多分类方法,构建模型集 M;

(3)将待测样本输入 SVM 多类分类器,如果待测

样本能够被 M 中某一模型识别,则认定有相应类别的流量被检测出。若被检出的流量为正常流量则继续,为异常流量则触发告警。循环执行(3)。如果待检测样本被判定为未知流量,则执行(4);

(4)将未知流量加入到待验证集合 P 中。定期对 P 中的流量进行分析,如果 P 中的流量可以聚类,并且与正常流量差异明显,则可认为出现了一种新的异常;

(5)将新的异常加入到(1)中的训练样本进行重新训练,获得新的模型集 M',用 M' 代替 M,执行(3)。

该方法利用改进的一对其余 SVM 多分类器进行检测并引入未知流量重训练机制,避免了传统"一对其余"方法中的分类重叠问题和样本不可分问题,缩短了检测时间,提高了检测效率。

3 异常流量检测实验

3.1 实验环境

为了测试该方法在实际系统中的检测效果,将其应用于如图 4 所示的异常检测系统中。该系统由采集模块、数据存储模块、异常流量检测模块和用户信誉管理模块四部分组成,其中,异常流量检测模块采用了文中提出的基于信息熵和改进 SVM 多分类器的异常流量检测方法。

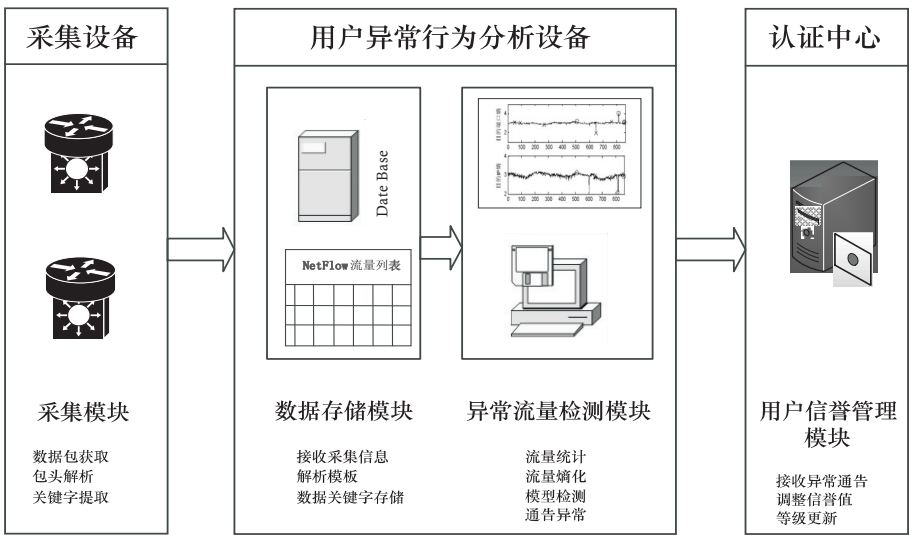


图 4 异常检测系统架构图

3.2 实验过程

实验中,通过采集校园网络出口链路数据包的方式获取正常流量。利用网络攻击模拟软件生成端口扫描、网络扫描、DoS 和 DDoS 等几种常见的攻击流量。在不同的检测周期内,开始时都只输入正常流量,然后在固定的时间段分别按 10%、20%、30% 和 40% 的比例混入攻击流量,获得异常流量。分别采用传统一对其余 SVM 多分类器和文中提出的改进的一对其余 SVM 多分类器在相同条件下进行实验,在异常流量检测模块的输出端观察检测结果。

3.3 实验结果

通过检测精度、误报率、检测率和检出时间四个指标,判断文中提出的异常流量检测方法能否满足实际检测系统的需要。各个检测指标的定义如下:

检测精度代表被正确分类的样本数占样本总数的比例;

误报率代表正常样本被误报为异常的个数占正常样本总数的比例;

检测率代表被检测出的异常样本数占异常样本总数的比例;

系统检测时间代表流量从输入采集模块开始,先后经过关键字提取、存储、熵值量化和检测,到获得最终检测结果所需的时间。

实验结果如表 1、表 2 所示。

表 1 传统一对其余 SVM 多分类器的检测结果

攻击流量占比/%	检测精度/%	误报率/%	检测率/%	系统检测时间/s
10	93.0	1.0	87.0	8.6
20	95.2	0.8	91.2	8.5
30	96.9	0.6	94.4	8.4
40	97.8	0.4	96.0	8.4

表 2 改进的一对其余 SVM 多分类器的检测结果

攻击流量占比/%	检测精度/%	误报率/%	检测率/%	系统检测时间/s
10	92.0	2.0	86.0	7.9
20	94.4	1.6	90.4	7.8
30	96.5	0.8	93.8	7.8
40	97.8	0.4	96.0	7.7

通过表 1 和表 2 发现,当攻击流量在异常流量中占比较小时,传统的一对其余 SVM 多分类器比提出的改进的一对其余 SVM 多分类器的检测精度和检测率

略高一些。随着攻击流量占比的增加,两种分类器的检测效果逐渐接近。当异常流量占比达到 40% 时,改进的一对其余 SVM 多分类器达到与传统分类器一样的检测精度和检测率。对比两种分类器的系统检测时间发现,改进的一对其余 SVM 多分类器所需的系统检测时间更短,检测速度更快。

4　结束语

文中结合信息熵理论和 SVM 分类算法,提出一种基于熵和改进 SVM 多分类器的异常流量检测方法。通过对不同流量在不同属性上的熵值进行分析,验证了信息熵对正常流量和异常流量在不同属性的特征分布上的差异性表达能力。针对传统“一对其余”方法可能遇到的分类重叠和不可分问题,提出一种改进的 SVM 多分类方法。将该方法应用于实际的异常检测系统中,通过实验证明了该方法对网络中常见的异常流量和攻击具有很好的检测效果。

参考文献:

[1] Nychis G,Sekar V,Andersen D G,et al. An empirical evaluation of entropy-based traffic anomaly detection[C]//Proc of Internet measurement conference. [s. l.]:[s. n.],2008.

[2] 朱应武,杨家海,张金祥. 基于流量信息结构的异常检测[J]. 软件学报,2010,21(10):2573-2583.

[3] 王海龙,杨岳湘. 基于信息熵的大规模网络流量异常检测[J]. 计算机工程,2007,33(18):130-133.

[4] 钱亚冠,关晓惠,王　滨. 基于最大信息熵模型的异常流量分类方法[J]. 计算机应用研究,2012,29(3):1019-1023.

[5] 吴　震,刘兴彬,童晓民. 基于信息熵的流量识别方法[J]. 计算机工程,2009,35(20):115-116.

[6] 范晓诗,李成海. 加权条件熵在异常检测中的应用[J]. 计算机应用研究,2014,31(1):203-205.

[7] 陈小辉. 基于数据挖掘算法的入侵检测方法[J]. 计算机工程,2010,36(17):72-73.

[8] 杜　强,孙　敏. 基于改进聚类分析算法的入侵检测系统研究[J]. 计算机工程与应用,2011,47(11):106-108.

[9] Denning D E. An intrusion-detection model[J]. IEEE Transactions on Software Engineering,1987,13(2):222-232.

[10] 陈锶奇,王　娟. 基于信息熵理论的教育网异常流量发现[J]. 计算机应用研究,2010,27(4):1434-1436.

[11] Lakhina A,Crovella M,Diot C. Mining anomalies using traffic feature distributions[C]//Proc of ACM SIGCOMM. [s. l.]: ACM,2005:217-228.

[12] Hsu C W,Lin C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks,2002,13(2):415-425.

[13] Song X. Multi-class classifier based on support vector machine and decision tree[J]. Computer Engineering,2005,31(14): 174-175.

[14] 焦春鹏. 基于二分类 SVM 的多分类方法比较研究[D]. 西安:西安电子科技大学,2011.

[15] 谭爱平,陈　浩,吴伯桥. 基于 SVM 的网络入侵检测集成学习算法[J]. 计算机科学,2014,41(2):197-200.

(上接第 30 页)

[3] 方金城,张岐山. 物流配送车辆路径问题(VRP)算法综述[J]. 沈阳工程学院学报:自然科学版,2006,2(4):357-360.

[4] 赵燕伟,张景玲,王万良. 物流配送的车辆路径优化方法[M]. 北京:科学出版社,2014.

[5] Laporte G,Mercure H,Nobert Y. An exact algorithm for the asymmetrical capacitated vehicle routing problem[J]. Network, 1986,16:33-46.

[6] 王德东,郑丕谔. 车辆路径问题的混沌神经网络解法[J]. 计算机集成制造系统,2005,11(12):1747-1750.

[7] Bullnheimer B,Hartl R F,Strauss C. An improved ant system algorithm for the vehicle routing problem[J]. Annals of Operations Research,1999,89(13):319-328.

[8] Dorigo M,Stützle T. Ant colony optimization:overview and recent advances[M]//International series in operations research & management science:handbook of metaheuristics. US:Springer,2010:227-263.

[9] 胡大伟,朱志强,胡　勇. 车辆路径问题的模拟退火算法[J]. 中国公路学报,2006,19(4):123-126.

[10] Barbarrosoglu G,Ozgur D. A tabu search algorithm for the vehicle routing problem[J]. Computers & Operations Research, 1999,26(3):255-270.

[11] 李　琳,刘士新,唐加福. 改进的蚁群算法求解带时间窗的车辆路径问题[J]. 控制与决策,2010,25(9):1379-1383.

[12] Chen C,Ting C. An improved ant colony system algorithm for the vehicle routing problem[J]. Journal of the Chinese Institute of Industrial Engineers,2006,23(2):115-126.

[13] Gan R,Guo Q,Chang H,et al. Improved ant colony optimization algorithm for the traveling salesman problems[J]. Journal of Systems Engineering and Electronics,2010,21(2):329-333.

[14] 刘晓勇,付　辉. 基于启发式蚁群算法的 VRP 问题研究[J]. 计算机工程与应用,2011,47(32):246-248.

[15] 张　锦,李　伟,费　腾. 交叉变异蚁群算法在 VRP 问题中的应用研究[J]. 计算机工程与应用,2009,45(34):201-203.

[16] Stützle T,Hoos H H. Max-min ant system[J]. Future Generation Computer Systems,2000,16(19):889-914.

[17] 耿耀华. 基于 MMAS 的配送线路规划研究与应用[D]. 济南:山东大学,2008.

基于熵和SVM多分类器的异常流量检测方法

作者： 朱佳佳， 陈佳， ZHU Jia-jia， CHEN Jia
作者单位： 北京交通大学 电子信息工程学院,北京,100044
刊名： 计算机技术与发展 ISTIC
英文刊名：
年，卷(期)： 2016,26(3)